IPRPD
International Journal of Arts, Humanities and Social Sciences
ISSN 2693-2547 (Print), 2693-2555 (Online)
Volume 06; Special Issue: August, 2025
DOI: 10.56734/ijahss.v6nSa2

REFRAMING AI IN EDUCATION: A VYGOSTSKIAN PERSPECTIVE ON PROXIMAL DEVELOPMENT

Mazzoni Elvis[1], Benvenuti Martina[2], Benassi Mariagrazia[3]

[1 2 3]*Department of Psychology University of Bologna, Italy*

Abstract

This paper explores the role of Artificial Intelligence (AI) in educational and clinical contexts through the lens of Vygotsky's cultural-historical theory. AI is conceptualized as a cultural and cognitive tool that can operate within the Zone of Proximal Development (ZPD), supporting individual learning and cognitive growth as well as societal changes. When functionally integrated into human practices, AI can become a "functional organ" that enhances human abilities—like how other tools (like a pen) extend motor functions. The focus is on active and critical use of AI and the role of human, which should foster reflective thinking, inner speech, and metacognitive regulation, especially through meaningful errors and adaptive feedback. Real-world applications such as *Proffilo* and *MATHia* are presented, along with a discussion of ethical concerns related to transparency, user agency, and personalized adaptation. The paper offers a theoretical framework for the ethical and developmental use of AI, highlighting the importance of co-design and empirical testing in real-life settings.

Keywords

Artificial Intelligence, Cognitive Growth, Proximal Development

1. Introduction

Across cultures, considerable efforts are being made to understand how to integrate artificial intelligence (AI) into educational settings. The main reason is that, although AI wasn't designed for educational purposes, it has become a pervasive tool in human everyday life without, however, operating within a clear educational framework.

Several perspectives have emerged so far to define AI's role in educational and learning contexts. Some authors highlighted its significant usage in defining personalized learning and educational paths to achieve individual objectives more quickly and efficiently than those gained by interacting with humans, thanks to AI capacity to handle and elaborate a huge amount of interconnected data. These findings have shocked those involved in education and clinical practice, who had previously ignored AI potential and underestimated its impact. However, they have also forced them to make important and necessary ethical and pragmatic considerations.

Another perspective showed that Artificial Intelligence in educational contexts can be considered as a tutor or supervisor able to act within the zone of proximal development (ZPD) of the learner. This space is defined by Vygostky as the difference between the "actual developmental level as determined by independent problem solving" and the "potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978; Cong-Lem and Daneshfar, 2024). From this point of view, AI can be seen as a functional organ (Leont'ev, 1972) that integrates a technological artifact (the AI system) with the cognitive human skills to achieve results otherwise impossible to realize separately. Like a pen that allows humans to write thanks to human motor skills, an AI system, functionally integrated into humans' skills and practices, would allow them to achieve a result higher than the actual development level. Following this perspective, an important aspect to consider is what does "functionally integrated" means? In a developmental perspective, an artifact is functional to human skills when it supports their potential, acting within the zone proximal development, without taking their place. A learner could use an AI system to translate a text into a foreign language. However, in this way it rests into her/his comfort zone of actual development. But she/he could go over the actual development, trying to translate the text into the foreign language, and using the AI system to supervise the result in the mother language, adapting it slavishly until the desired result is achieved. In this second case, she/he acts within her/his zone of proximal development, improving her/his competences. This perspective opens relevant and interesting developments to design educational and learning activities in which AI systems could play an important role in improving learners' skills acting into their

zone of proximal development, based on their actual developmental level and recognizing their special needs. Relevant methodological questions remain open about the ethical issues that must guide the development of AI algorithms acting as functional organs for development and determine human-AI interaction.

2. A new framework for AI: the role of Vygotsky's theory

When adopting the Vygotskian perspective (Cong-Lem and Daneshfar, 2024), we consider AI as a tool that could change human and societal development. Indeed, its usage would change the social and cultural process, determining individual maturation, in terms of cognitive and social competencies. The Vygotskian approach could be used as a framework to re-define the educational and psychological perspectives about human-AI interaction in educational or clinical settings. In this regard, the present paper addresses a critical gap in current AI-in-education literature by reframing AI through a Vygotskian developmental lens, in contrast to dominant behaviorist and data-driven paradigms. Traditional behaviorist models view AI as a dispenser of stimuli, rewards, and corrective feedback, aiming to condition learner responses through reinforcement. This perspective lacks the cultural and environmental factors interacting within the system and oversimplifies the AI effects on societal level and ignores the cultural and ethical consequences of AI misuse. Similarly, data-driven systems emphasize performance tracking, predictive analytics, and automated adaptation, often reducing learning to measurable outputs. Within this perspective, Intelligence Tutoring Systems, which aims to provide immediate and personalized feedback to students, are sometimes used only for helping students learning specific topics, according to individual's objectives and learning outcomes, but fail to educate about "how to" learn and what are the best strategies to use within specific contexts.

In contrast, this work positions AI as a possible artifact or "functional organ" embedded within the learner's Zone of Proximal Development (ZPD), where cognitive growth is supported through social interaction, internalization, and scaffolded reflection. Drawing from Vygotsky's cultural-historical theory, the paper emphasizes the importance of inner speech, symbolic mediation, and meaning-making—dimensions largely neglected in behaviorist or technocratic models. Furthermore, by integrating ethical considerations into this developmental framework, this approach offers a novel, human-centered approach for AI algorithm design that prioritizes learner agency, transparency, and co-construction within cultural framework. Within this framework, the human cultural role is crucial in interacting with AI and in consciously determining societal challenges: humans have a pro-active role in AI development, usage and education, able to learn and enable them to take initiative and anticipate future human-machine interactions, rather than simply responding to stimulation according to predetermined non ethical algorithms. Essentially, in Vygostkian perspective, differently from data-driven-learning or behaviorist models views, interaction with AI is driven by human culture, who should be prepared (e.g. educated) to take care of the issues related to interaction with AI tools itself. This positions the perspective as theoretically grounded and practically significant contribution to current debates in AI and education.

3. AI Language, symbols and artifacts

In Vygotskian theoretical framework, AI could be considered an artifact to improving human language and symbols.

Language, symbols, and artifacts are the conceptualization of learning tools that do not simply transmit knowledge as interacting with environment, but they shape cognition itself, transforming basic mental functions into higher-order thinking, deepening the learning process. Language is not just a channel of communication with others; it is the way the child should develop thought and learn. By language development, everyone is enabled to internalize external actions, experiences done in collaboration with peers: what is first done socially (e.g., problem-solving with a peer) becomes a mental process through internal speech. Accordingly, through language, the child learns how to regulate themselves, organize and plan activities, and improve critical thinking and abstract thinking. For Vygotsky, this is the shift from *external speech* to *inner speech* and marks a key developmental step in cognitive growth. Within the language, a particular tool for development consists of the symbols that are culturally created systems of meaning (e.g. written language, numbers, mathematical signs, musical notation, gestures). They represent knowledge and abstract concepts and during development they change according to cultural rules and family traditions. Mastery of symbolic systems allows individuals to manipulate ideas, communicate complex thoughts, and solve problems within and outside their environment. While language and symbols are given as internal tool, artifacts (both physical and cognitive) are man-made tools or instruments. They enable learners to externalize, organize, and extend their mental processes. Vygotsky referred to tools as "mediators" between human action and the environment. Mastery of tools reshapes cognition improving development, while nonfunctional tool usage constitutes an obstacle for development, contributing to mental health issues.

What is the role of AI in rebuilding human language and communication process? Referring to AI, as in the form of NLP (Natural Language Processing algorithms), the AI producing language could be potentially used as a tool for simulating communication and thinking (favoring the inner speech) and a tool for sociality (given by spoken or written interactions). Therefore, AI could contribute to development, whenever it is possible (i.e. it has been developed for), for inducing critical thinking. Not AI for producing responses but for stimulating questions and enhance human-to-human interaction. The system could incorporate this feature by creating it based on specific

requirements, allowing users to ask questions, request solutions as well as provide feedback, and supporting active user participation in the communication process. All these features could be interconnected within social and cultural activities and embody cultural values and cognitive strategies, whenever the algorithm is created accordingly. Within this type of space, teachers, peers and AI human tutors are determinant for the functional AI usage in developmental contexts, such as in school and in clinical settings. The developmental process embedded within the algorithm should consider also how to elicit creativity and the possibility to enhance learning by errors. Creativity allows learners to go beyond what is immediately available in the environment—to recombine, transform, and reimagine existing knowledge about the world, in different ways. Learning process is possible through the error meaningful, conducting to learning awareness and motivation. Rooted in constructivist and metacognitive theories, this perspective posits that errors are not simply failures but powerful opportunities for conceptual change, self-regulation, and the expansion of cognitive boundaries. When framed appropriately, mistakes activate reflection, highlight knowledge gaps, and initiate corrective strategies—all central to meaningful learning. Errors often occur at the boundary of ZPD - where cognitive challenge meets insufficient mastery. These moments of impasse are not accidental: they are markers of cognitive disequilibrium, ideal for intervention by a more knowledgeable other, be it a teacher, peer, or also the intelligent system.

AI can be understood as a complex cultural artifact, because it is a technological tool created by our society to extend human capabilities and overlap several cognitive issues. In this sense, as a tool could be functional when it contributes to positive development or is dysfunctional when it does not allow individual and societal development.
AI could be considered an effective artifact when:

a. it mediates learning by adaptive feedback, given with personalized instruction for problem solving and task management, or improving language processing, or real-time scaffolding.
b. It can extend cognitive activity, allowing learners to perform tasks they could not achieve unaided (e.g., via intelligent tutoring systems, Natural Language Processing tools, or adaptive videogames).
c. It shapes how learners interact with content, solve problems, and even understand themselves (through the possibility of creating suggestions for data reflection and metacognitive insights).
d. Include learning by error paradigm, by assessing and monitoring errors not just for scoring, but as diagnostic signals of a learner's current developmental level. AI systems can then modulate the intensity, type, and frequency of scaffolding to maintain the learner within their learning zone, enabling productive struggle rather than frustration.

A research-based example of AI tool integrated in the learning process is Proffilo (Orsoni et al., 2022) a serious game based on machine learning aimed to assess and classify learner's cognitive profile taking into account several cognitive abilities (logic, memory, attention, visual perception, phonological awareness, verbal comprehension). In Proffilo, the AI algorithm is specifically designed to allow teachers and students to discover their cognitive potential and fragility and to find possible clusters characterizing neurodiversity. Proffilo has been co-designed with students and teachers trying to enhance the ability of the algorithm in classifying multiple types of cognitive profiles that could be partially overlapping and that, in that way, represent the richness manifestation of neurodiversity. Indeed, the term "neurodiversity" represents the concept that there is natural variation in how people's brains work, and that the algorithm should capture not only deficits related to neurodevelopmental disorders, but also potential and similarities among cognitive profiles that capture their cognitive heterogeneity. In this way, the algorithm's development was designed not only to consider the expected final results of AI clustering, in terms of grouping people with similar cognitive profiles, but also to mitigate the risks of not properly representing neurodiversity, including avoiding stigma and oversimplification of individual characteristics, as well as potential errors and biases due to the algorithm. To this aim, the AI outcome responds primarily to teachers and students needs, taking into account that learning difficulties or emotional issues could determine difficulties in recognizing cognitive functioning potential. Although in the general population cognitive functions are clearly linked and specifically correlated with learning outcomes (e.g. logic and visual perception are correlated with math abilities), some subgroups of individuals could not be represented by such correlations. At individual level, cognitive heterogeneity could manifest very differently, not following rules about the general population and AI could help disentangle this heterogeneity. The AI system in Proffilo is a machine learning classifier that finds possible cognitive clusters resulting from combination of different cognitive abilities. This clusters are aimed to support education by giving new perspective of interpreting learners' diversity in cognitive profile and help teachers to reframe new didactic strategies taking into account cognitive potential as well as fragilities. This perspective in the case of neurodiversity is particularly challenging because didactic are usually tailored on deficits and not on potentials. Another application is *Carnegie Learning's MATHia (Pane et al., 2014; Ritter et al., 2017)*. MATHia is an AI-driven intelligent tutoring system (ITS) for improving math learning in middle and high school using gamified exercises. MATHia monitors the student's problem-solving behavior in real time, offering context-specific hints and step-by-step guidance, and uses knowledge tracing to estimate a learner's mastery level and adapt instruction accordingly. In this sense, it constantly monitors the individual responses within the ZPD and reframing the context accordingly. However, both Proffilo and

MATHia systems do not give static responses to the learner's demand but poses possible responses and interpretations that should be considered as inputs to teachers and students enabling open discussions about how to improve learning strategies and continue creating learning opportunities. In other words, they are created as tools that should be used as learning facilitators only when they are guided correctly by humans and not to be used without human expert control. The validation of Proffilo and MATHia should be assessd not only by measuring the accuracy of the classification or the improvement in math abilities but also on the users experience: in Proffilo, the AI effectiveness is given when the results produced by the algorithm are correctly perceived and used by students and teachers to increase neurodiversity awareness; the effectiveness of MATHia is realized when the improvement obtained by AI training are generalizable to other contexts outside of the game experience and the strategies learnt by MATHia could be extensively used in similar math exercises and could be discussed in different settings (i.e. at school).

4. The Zone of Proximal Development and its connection with AI

In Vygotskian perspective, the AI as a tool should be considered within a specific space of the human developmental experience, coexisting with educators and not replacing them. In fact, AI can give a picture of the individual's zone of Proximal Development (ZPD) taking into account multiple information (Holmes et al., 2018) and help educators to integrate different perspective about individual's learning potential.

Following this perspective, it is possible to say that AI aligns with the ZPD in the following modes:

a. Adaptive Scaffolding: AI systems can provide personalized hints, feedback, or challenges based on the learner's current level, helping them progress just beyond what they can do alone.
b. Real-time Support: Through technologies like machine learning, AI can continuously assess a learner's performance and adjust support dynamically—just like a good tutor would.
c. Encouraging Active Engagement: Rather than doing tasks for the learner, effective AI tools work with the learner—promoting exploration, reflection, and internalization of new skills.
d. Monitoring and Feedback on Errors: AI can interpret mistakes not as failures, but as diagnostic indicators of developmental stage, tailoring feedback that keeps the learner in the ZPD rather than letting them feel frustrated or unchallenged.
e. Functional Integration: when AI becomes a "functional organ" (Leont'ev), it's no longer just a tool – it is embedded in the learner's cognitive process, helping to extend their learning capabilities beyond their actual level.

This last point is relevant because it offers food for thought on how AI can be considered a "functional organ" in the sense of a new tool enabling to change our way to learn and educate.

5. AI as a "Functional Organ"

Vygotsky's collaborator, Leont'ev introduced the notion of "functional organs"—when an artifact (e.g., a pen, a notebook) becomes so integrated into our thinking that it essentially becomes part of our cognitive system. A well-designed AI system for learning becomes a functional organ that enables the learner to extend beyond their actual level of development. For example, a child using a gamified AI system to solve math problems with intelligent hints and visual modelling is engaging with a system that becomes part of their cognitive process, not just an external tool. Indeed, when AI is used as a functional organ, it can serve as the initial external scaffold for attention regulation, strategy use, or decision-making. With repeated use and reflection, the learner may appropriate these processes, internalizing the methods. So, AI doesn't just support learning, it can become part of how the learner thinks, solves problems, and creates meaning.

This concept of AI, interacting with humans as a functional organ, has analogies with that evoked by Floridi, considering AI as artificial agent enables us to solicit responses and dialogue and recalls the need of ethical considerations about how to create AI algorithms that, by design, are guided by humans and could contribute to educate humans about human-AI interaction. This point is linked to that discussed by Floridi (2025), who suggested reframing AI as an Artificial Agent, created by humans for humans and moved beyond biological and anthropomorphic misconceptions of AI, recognizing for its unique characteristics: being a functional tool embedded in human thoughts. This conceptual shift offers a more robust basis for analyzing both the challenges and opportunities presented by AI in education and psychological fields, while also supporting more informed discussions about their future development and broader societal implications, centered on ethical principles that anticipate technological development, and not *viceversa*. In this perspective, artificial agency is not opposed to human agency but could be conceptualized as a collaborator to enhance human capabilities while maintaining ethical standards of responsibility and accountability (Langley et al., 2017). Accordingly, considering AI in Vygotskian view, the most valuable AI systems are those that provoke cultural and societal reflection, dialogue, and internalization, rather than replacing the thinking process.

6. Ethical Considerations

Within the framework of the study of human development, human-AI interaction involves broad ethical issues related to the concept of space of development and agency that delimits the workflow of AI in interacting with humans (Jobin, Ienca, & Vayena, 2019). If we adopt the perspective in line with Floridi (2025), according to which AI could be considered an "artificial agent", then we would accept that even if AI is developed and driven by humans for human interactions, it possesses three main characteristics that define it as an agent: interactivity, autonomy and adaptability. Interactivity indicates the capability to act on the environment and to be acted upon by it. Autonomy is the ability to initialize state changes independently of external action. Adaptability is the capability to change behavior in response to stimuli. However, the nature of AI agency is still critical and has still been defined appropriately taking into consideration that AI agency is relying on human interactions (with developers, owner, users and stakeholders). AI agency is not a "full" or "real" agency, it is only a "functional" agent, which exist upon human agency and lacks ontology. Ontologies provide a formal and computable representation of knowledge within a domain, defining not just *what* entities exist, and their relations but also the *how* and the *why* they relate upon certain scopes and objectives, justifying interactions, and behavior within a given system, guided by attitudes and perspective towards future. Indeed, ontology consists of a set of concepts, relationships, and constraints that enable semantic interoperability and automated reasoning within a cultural and historical framework. This capability is especially critical in high-demanding domains where detecting complex interactions is crucial for understanding and guiding intentional interactions (such as in educational and clinical settings). While algorithms are enabled to use taxonomies, these could be only soiled and careless because of their artificial nature. Indeed, AI foundational scopes are usually courtesy hidden and are not explicitly stated by AI developers nor by AI owners. This raises ethical issues, especially when humans are not experts, or minors or people with mental health conditions (Drigas and Ioannidou, 2013; Castellani et al, 2023; Barua, P.D., et al., 2022). So how do we enable people with inherently limited expertise in AI processes to recognize the limits and possible bias of AI? AI bias and errors constitute one of the main issue related to AI application in education (Baker and Hawn, 2022). Different types of bias emerged from literature: statistical biases related to the type of algorithm used and bias of measurement and error related to imbalances in how well the model performs across groups, to disparate impacts and discrimination as different interpretations are applied and controlled. The ways to measure bias and minimize them are multiple and are related to awareness about AI limitation and explainability. The ethical questions should anticipate the AI development: Is it possible to make the goals of AI transparent and explicit in these cases? If it is not possible for the user, then we should have a human expert (such as a teacher or clinician) who can act as a tutor in the human-AI relationship. Transparency is one of the ethical principles guiding the scope of ethical AI-based products and it is particularly urgent. One way to achieve it can be co-designing with users, enabling them to take part in AI development and give their own ontological view in the human-AI interaction. Building AI tools based on ethics means designing them starting from the ethical principles that we want to adopt (Ethic by design process). But we need to make sure that these ethical principles are discussed with all the users and stakeholders involved and made explicit so that those who use AI have still the freedom and agency during their use. In the case of absence of ethical guidance of AI or opacity in AI development (such in the case of lack of transparency and explicability), then it would be difficult for the user to have a functional use of AI in terms of functional organ. Because in this case the user and the AI owner or developer do not share their own ontology given different semantics to the interactions and to their final goals.

7. Limits and Critical Perspectives on AI's Pedagogical Role

While this paper advances a developmental and ethically grounded perspective on AI in education, it is equally important to acknowledge the limitations and risks associated with the pedagogical use of AI technologies—particularly when these systems are adopted without critical reflection. One of the foremost concerns is epistemic opacity. Many AI systems, especially those using deep neural networks, operate through complex algorithmic processes that are not transparent to users. Teachers and learners alike may receive personalized feedback or instructional interventions without understanding how those recommendations were generated or what assumptions underline them (Burrell, 2016; Jobin, Ienca, & Vayena, 2019). This undermines the principle of epistemic agency, a foundational element in educational contexts where learners should be encouraged to evaluate knowledge claims, ask questions, and understand the reasoning behind feedback. When AI feedback becomes authoritative but inscrutable, there is a risk that learners may begin to accept machine-generated guidance uncritically, weakening their ability to engage in reflective thinking and conceptual justification (Holmes et al., 2018). Moreover, the over-reliance on AI for cognitive support may inadvertently lead to a kind of learned helplessness, especially in younger or neurodiverse learners. Instead of challenging students within their Zone of Proximal Development (ZPD), some AI systems may offer immediate solutions or adapt tasks to such an extent that learners are not required to struggle productively (Castellani et al., 2023; Barua et al., 2022). While such systems may produce short-term gains in performance metrics, they may erode the deeper developmental goals of autonomy, self-regulation, and error-based learning that this paper has argued are central to the educational process. In parallel, there is a growing concern about the deskilling of

educators. As AI systems increasingly take on roles traditionally occupied by teachers—such as identifying learning difficulties, adjusting task difficulty, or providing motivational feedback, there is a danger that teachers may be marginalized as passive implementers rather than active shapers of the learning process (Selwyn, 2019). This could not only diminish professional expertise but also devalue the emotional and relational dimensions of pedagogy, which are critical to inclusive, responsive education. Furthermore, many AI systems are designed around optimization goals that reflect economic or administrative priorities—efficiency, productivity, standardization—rather than educational values like curiosity, critical inquiry, or democratic participation (Williamson, 2017; Knox, 2020). These underlying logics risk shifting the goals of education itself, privileging measurable outcomes over developmental depth. Finally, there are pressing ethical and political questions about data ownership, consent, and surveillance, particularly in systems that continuously monitor learner behavior. When learners are profiled by opaque systems and assigned risk scores or learning paths without transparency or recourse, the potential for bias, misclassification, or harm becomes significant—especially for marginalized groups (Drigas & Ioannidou, 2013; Jobin et al., 2019). These critiques do not negate the developmental promise of AI; rather, they underscore the need for co-designed, transparent, and context-sensitive systems that foreground pedagogy over prediction and empowerment over automation. Any vision of AI as a functional organ must be tempered by a commitment to preserving human judgment, critical engagement, and the complexity of learning as a social, ethical, and developmental act (Floridi, 2025).

8. Conclusion and Future Directions

In this paper, we argued for fundamental reorientation in the way Artificial Intelligence is conceptualized, designed, and applied within educational and clinical contexts. Drawing on Vygotsky's cultural-historical theory and the concept of the Zone of Proximal Development (ZPD), we positioned AI not as an automated tutor or behaviorist feedback machine, but as a culturally embedded and developmentally potent tool—a functional organ that can mediate learning and support human cognitive growth when ethically and intentionally integrated. Where dominant models of AI in education tend to emphasize behaviorist principles (such as reinforcement, conditioning, and performance tracking) or rely heavily on data-driven personalization and algorithmic optimization, our framework challenges the reduction of learning to measurable outputs or behavioral performance. Instead, we advocate for a developmental view of AI—one that recognizes learners as active meaning-makers, embedded in cultural and social systems, whose development depends not just on information delivery or correction but on the appropriation of tools, language, and symbolic systems. From this standpoint, AI becomes a mediator within the ZPD: a dynamic space where learners move from actual to potential development through guided participation, reflection, and social interaction. Importantly, this mediation is not merely instrumental; it is transformative. When AI tools are designed to scaffold thinking, provoke inner speech, and support metacognitive regulation—rather than merely automate tasks or monitor errors, they align with the core processes of learning as envisioned by Vygotsky. Errors, in this framework, are not failures to be eliminated but opportunities for growth, cognitive disequilibrium, and reflection. AI can play a crucial role here, not by preventing mistakes, but by recognizing them as signals of developmental readiness and responding with context-sensitive scaffolding. The concept of AI as a functional organ, borrowed from Leont'ev, further extends this idea by emphasizing how artifacts—when fully integrated into cognitive activity—can become part of the learner's extended mind. Much like a pen enhances motor-cognitive coordination or language structures thought, a well-designed AI system can become a constitutive element of how learners think, plan, self-regulate, and create. This integration, however, is not automatic; it requires intentional design, pedagogical alignment, and above all, ethical transparency. Indeed, ethical considerations are not an appendix to this framework; they are intrinsic to it. Because AI systems operate within deeply human domains (development, learning, identity), their design must be guided by ethical principles that safeguard user agency, ensure interpretability, and respect cognitive diversity. Particularly in contexts involving minors, neurodivergent learners, or clinical interventions, the opacity or misalignment of AI systems can lead to misuse, disempowerment, or harm. We proposed that these challenges can be addressed by embracing co-design methodologies, involving users and stakeholders not just as recipients of AI but as active participants in shaping its ontological assumptions, goals, and interaction paradigms. Examples such as Proffilo and MATHia demonstrate the promise of AI tools designed with developmental sensitivity and practical application in mind. These systems do not claim to replace teachers or learners' cognitive work; rather, they support the interpretive process, offer adaptive feedback, and create space for collaborative exploration. Such tools can be especially empowering when they help reveal a learner's cognitive potential, often obscured by emotional or behavioral challenges—and foster new strategies for personalized support. Ultimately, our proposal is not simply to apply Vygotsky to AI, but to use his theory as a foundational framework for reimagining what it means to learn with and through AI tools where intelligence is human guided. This perspective reframes AI not as an external authority or solution provider but as a co-agent in human developmental tool whose function is to provoke reflection, support symbolic mediation, and expand the learner's capacity to act in the world. Future research must move toward empirical validation of this framework by testing how AI tools can be co-constructed, meaningfully integrated, and ethically governed in real-world educational and clinical settings. Interdisciplinary collaboration among educators, psychologists, developers, and learners will be essential to refine our understanding of functional integration,

scaffolded agencies, and ethical co-design. Only through such dialogue can we ensure that AI serves not only educational outcomes but the broader goals of human flourishing, creativity, and developmental empowerment. In sum, the developmental, cultural-historical perspective we present offers a timely and much-needed alternative to prevailing technocratic paradigms. It invites a more nuanced and humane approach to AI in education—one that respects the complexity of human thought, the social nature of learning, and the ethical imperatives of responsible technological innovation.

9. Recommendations for educators or developers

To translate these ethical principles into practice, educators and AI developers should adopt a co-design approach that actively involves teachers, students, and other stakeholders in shaping the purpose, function, and feedback mechanisms of educational AI tools. Developers should prioritize transparency by ensuring that systems provide interpretable outputs and explanations for their recommendations, especially in contexts involving minors or neurodiverse learners. Educators, in turn, should be equipped with AI literacy training that enables them to critically assess how AI systems support—or constrain—developmental goals such as autonomy, metacognition, and meaningful error-making. Both groups should avoid framing AI as a replacement for pedagogical judgment and instead treat it as a scaffold for learner agency, carefully calibrated to the learner's developmental level. Importantly, systems must be designed with ethical safeguards that protect user privacy and allow opt-in control over data use, while allowing teachers to intervene, question, or override AI decisions when needed. Embedding these practices supports the creation of AI that functions not as an opaque authority but as a collaborative, developmentally aligned partner in the learning process.

References

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International journal of artificial intelligence in education, 32(4), 1052-1092*.

Barua, P. D., et al. (2022). Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders. *International Journal of Environmental Research and Public Health*, 19(1192). https://doi.org/10.3390/ijerph19031192

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). https://doi.org/10.1177/2053951715622512

Castellani, A., et al. (2023). Ethical artificial intelligence in telerehabilitation of neurodevelopmental disorders: a position paper. In *International Conference on Computational Science and Its Applications* (pp. 87-103). Cham: Springer Nature Switzerland.

Cong-Lem, N., & Daneshfar, S. (2024). Generative AI and Second/Foreign Language Education from Vygotsky's Cultural-Historical Perspective. In Bui, H. P., & Namaziandost, E. (Eds.), *Innovations in Technologies for Language Teaching and Learning* (Studies in Computational Intelligence, Vol. 1159). Springer, Cham. https://doi.org/10.1007/978-3-031-63447-5_10

Drigas, A., & Ioannidou, R. E. (2013). A Review on Artificial Intelligence in Special Education. Springer.

Floridi, L. (2025). Artificial Intelligence as a New Form of Agency (not Intelligence) and the Multiple Realisability of Agency Thesis. Philosophy & Technology 38:30. https://doi.org/10.1007/s13347-025-00858-9

Holmes, W., Dragon, T., & Jiménez, J. (2018). *Adaptive scaffolding for concept learning in intelligent tutoring systems: iTalk2Learn case study*. *Computers & Education*, 126, 178–191. https://doi.org/10.1016/j.compedu.2018.07.001

Jobin, A., Ienca, M., & Vayena, E. (2019). *The global landscape of AI ethics guidelines. Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017, February). Explainable agency for intelligent autonomous systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*(Vol. 31, No. 2, pp. 4762-4763).

Leont'ev, A. N. (1972). *Activity, Consciousness, and Personality*. Prentice-Hall.

Orsoni, M., Giovagnoli, S., Garofalo, S., Magri, S., Benvenuti, M., Mazzoni, E., & Benassi, M. (2023). Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile. *Heliyon*,*9*(3).

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, 36(2), 127–144.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2017). *Cognitive Tutor: Applied research in mathematics education*. *Psychonomic Bulletin & Review*, 24(2), 597–603. https://doi.org/10.3758/s13423-016-1117-3

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*. SAGE.